

This Page Is Inserted by IFW Operations  
and is not a part of the Official Record

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,  
please do not report the images to the  
Image Problems Mailbox.**

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-062693

(43)Date of publication of application : 07.03.1997

(51)Int.Cl.

G06F 17/30

(21)Application number : 07-215670

(71)Applicant : HITACHI LTD

(22)Date of filing : 24.08.1995

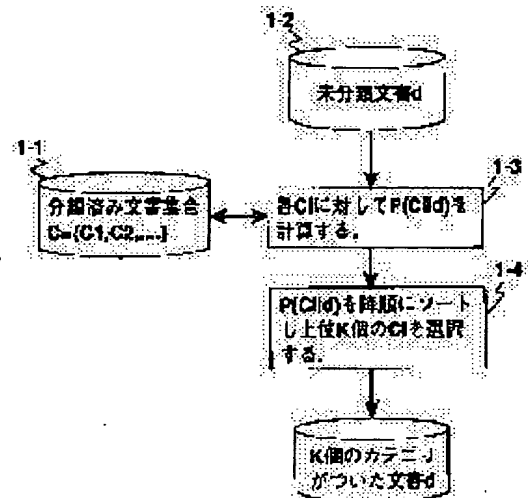
(72)Inventor : IWAYAMA MAKOTO  
MOTODA HIROSHI

## (54) DOCUMENT CLASSIFYING METHOD ACCORDING TO PROBABILITY MODEL

## (57)Abstract:

PROBLEM TO BE SOLVED: To automatically classify a document without interpolating any data even when features are applied by a keyword not existent in data for exercise by expressing the document in the set of words and applying the features to the document corresponding to the event of probability for extracting certain words from that word set at random.

SOLUTION: A non-classified document 1-2 is classified by using an already classified document set 1-1. In this case, the classified document set 1-1 is expressed in  $C=[C1, C2, \dots]$ . At a processing function 1-3, the 'probability  $P(Ci \text{ versus } d)$  for a non-classified document (d) to be classified in a category  $Ci$ ' is calculated. Namely, the 'probability  $P(Ci \text{ versus } d)$  for the non-classified document (d) to be contained in the document set already classified as  $Ci$ ' is calculated. At a processing function 1-4, the  $P(Ci \text{ versus } d)$  calculated by the processing function 1-3 is sorted in the descending order and K pieces of high-order categories are selected, for example, and defined as categories provided in the document (d).



## LEGAL STATUS

[Date of request for examination]

02.11.2000

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2000 Japanese Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平9-62693

(43) 公開日 平成9年(1997)3月7日

(51) IntCl.<sup>6</sup>

G 0 6 F 17/30

識別記号

庁内整理番号

F I

G 0 6 F 15/401

技術表示箇所

3 1 0 A

3 1 0 D

審査請求 未請求 請求項の数1 O L (全 3 頁)

(21) 出願番号 特願平7-215670

(22) 出願日 平成7年(1995)8月24日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 岩山 真

埼玉県比企郡鳩山町赤沼2520番地 株式会  
社日立製作所基礎研究所内

(72) 発明者 元田 浩

埼玉県比企郡鳩山町赤沼2520番地 株式会  
社日立製作所基礎研究所内

(74) 代理人 弁理士 小川 勝男

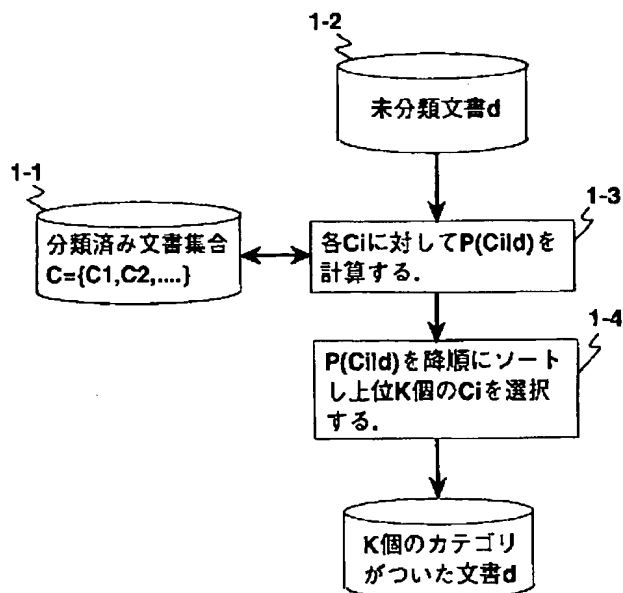
(54) 【発明の名称】 確率モデルによる文書分類方法

(57) 【要約】

【目的】 分類済みの文書集合（訓練用データ）をもとに新たな文書を分類する際、訓練用データが不十分な場合でもデータ補完を行うことなしに文書分類が可能になる文書の確率的特徴付けの方法とそれに基づく文書分類の方法の提供。

【構成】 特徴付けの対象となる文書集合Cは、それが含む単語の集合Wで表現される。ここで、単語集合Wから単語を無作為抽出する事象を考え、無作為抽出した単語がある特定の単語 $w_i$ と等しいという事象を $T=w_i$ とおき、確率 $P(T=w_i | C)$ を推定する。事象 $T=w_i$ は全ての $w_i$ に関して背反であるため、Wに含まれる全ての単語 $w_i$ に対して確率 $P(T=w_i | C)$ を推定し、それらを総和した確率は、文書集合Cを単語集合Wで特徴付けたことになる。

図1



## 【特許請求の範囲】

【請求項1】 訓練用データとしての分類済みの文書をもとに新たな文書を確率的に分類する際、文書を単語の集合として表現しておき、その集合から単語をランダムに抽出する確率事象により文書の特徴付ける方法。

## 【発明の詳細な説明】

## 【0001】

【産業上の利用分野】 本発明は、大量の文書を確率的に分類する手法に係わるものであり、従来は人手で行っていた文書分類を自動的に行う仕組みを与えるものである。

## 【0002】

【従来の技術】 確率的に文書分類を行うためには、文書をいかに特徴付けるかが問題になる。従来の手法は、「文書がある単語（キーワード）でインデックスされる／されない（文書にある単語が含まれる／含まれない）」という基本事象により文書の特徴付けを行っていた。例えば、文書の集合Cを単語wで特徴付けることを考えると、「文書集合Cの中からランダムに抽出した文書が単語wでインデックスされる確率」を $P(w=1|C)$ と書き、この確率により文書集合Cの特徴付けを行っていた。ここで、 $P(w=1|C)$ はCの中のk個の文書が単語wを含んでいる場合、「 $k/(Cに含まれる文書数)$ 」で推定できる。複数の単語集合 $W=\{w_1, w_2, \dots, w_m\}$ で文書集合Cを特徴付けるには、 $P(w_1=1|C)*P(w_2=1|C)*\dots *P(w_m=1|C)$  (1) を計算すればよい。

【0003】 文書分類では、分類に先だって分類済みの文書があらかじめ訓練用データとして与えられている。今、文書集合Cとして、訓練用データの中で同じカテゴリcに分類されている文書集合を設定すれば、前記(1)式の確率はカテゴリcを特徴付けることに相当し、この特徴付けを用いて新たな文書を分類することが可能になる。この際、これから分類しようとする文書dも同じく単語で特徴付けるのだが、これは前記(1)式において $C=\{d\}$ とした場合に相当する。

## 【0004】

【発明が解決しようとする課題】 上記従来技術の問題点は、もしCの中の文書どれにもある単語 $w_i$ を持っていなければ $P(w_i=1|C)$ が0になるため(1)の確率全体も0になってしまうことである。このような単語は文書の特徴付けとして使わないのが望ましいが、事前に単語 $w_i$ を特定することは難しい。そこで従来技術では、データ補完（スムージング）の手法により $P(w_i=1|C)$ が0にならないような補正を行う。ところが、補完の正当性を保証することは一般に困難である。

【0005】 本発明の目的は、上記の状況において、文書集合Cを特徴付ける全体の確率が上記のような単語 $w_i$ に影響されないような頑強な文書特徴付けの確率的方法を提供することである。

## 【0006】

【課題を解決するための手段】 上記目的は、文書を単語の集合で表現し、単語集合からある単語をランダムに抽出するという確率事象で文書の特徴付けることで達成される。

## 【0007】

【作用】 文書集合 $C=\{d_1, d_2, \dots, d_n\}$ を単語集合 $W=\{w_1, w_2, \dots, w_m\}$ で特徴付けることを考える。ここで、Cの各文書 $d_i$ は文書 $d_i$ に含まれるキーワードの集合で表現されている。例えば、文書 $d_i$ が単語 $w_1$ を1個、 $w_5$ を3個、 $w_8$ を2個含んでいれば、

$$d_i = \{w_1, w_5, w_5, w_5, w_8, w_8\} \quad (2)$$

となる。Cの表現形式は、Cに含まれる各々の文書表現形式（つまり上記の各集合 $d_i$ ）の和集合となる。

$$【0008】 C = d_1 \vee d_2 \vee \dots \vee d_n$$

（“ $\vee$ ”は和集合の演算子） (3)

ここで、単語集合からある単語を無作為抽出する事象を考え、無作為抽出した単語が $w_i$ と等しいという事象を $T=w_i$ と置く。この事象は全ての $w_i$ に関して背反であるため、各事象に対して総和した確率、

$$P(T=w_1|C)+P(T=w_2|C)+\dots +P(T=w_m|C) \quad (4)$$

は、文書集合Cを単語集合Wで特徴付けたことに相当する。 $P(T=w_i|C)$ は、Cに $w_i$ がk個含まれている場合、「 $k/(Cに含まれる単語数)$ 」で推定できる。(4)式において、全体の確率が各単語に関する確率の和の形になっていることに注意されたい。

【0009】 従来の確率(1)は積の形になっている。よって、Cのどの文書も単語 $w_i$ を持っていない場合を考えると、 $P(T=w_i|C)$ は0になるが、和形式としたことにより全体の確率(4)も0になってしまうことはない。従来の確率(1)では積形のため全体の確率が0になってしまうことが問題であった。

## 【0010】

【実施例】 以下、本発明の実施例である自動文書分類について説明する。

【0011】 図1に自動文書分類の概要を示す。自動文書分類では、既に分類済みの文書集合1-1を用いて、未分類の文書1-2を分類する。ここで、分類済み文書集合1-1は、 $C=\{C_1, C_2, \dots\}$ と表現される。各 $C_i$ はカテゴリ $C_i$ と分類されている文書の集合である。よって、例えば、ある文書がカテゴリ $C_1$ とカテゴリ $C_3$ に分類されている場合、その文書は文書集合 $C_1, C_3$ 両方に含まれることになる。

【0012】 処理機能1-3においては、「未分類の文書dがカテゴリ $C_i$ に分類される確率 $P(C_i|d)$ 」を計算する。言い替えると、「未分類の文書dが、既に“ $C_i$ ”として分類されている文書集合に含まれる確率 $P(C_i|d)$ 」を計算することになる。

【0013】 この確率 $P(C_i|d)$ を計算するため

に、本発明で提案した文書の特徴付けを用いる。具体的には、「ある単語集合から無作為に抽出したキーワードが  $w_j$  と等しい」という事象  $T=w_j$  を考える。この事象は、全ての単語に関し背反であるため、各事象について  $P(C_i | d)$  を条件付けると、

$$P(C_i | d) = P(C) * \sum_j [P(T=w_j | C) * P(T=w_j | d) / P(T=w_j)] \quad (6)$$

となる。(6)式の

$$\sum_j [P(T=w_j | C) * P(T=w_j | d)]$$

は、文書  $d$  とカテゴリ  $C_i$  を単語集合で同時に特徴付けたことに相当している。

【0014】各確率は以下の方法で推定できる。

【0015】 $P(T=w_j | C_i) =$  「 $C_i$  に含まれる単語  $w_j$  の数 /  $C_i$  に含まれる単語数」

$P(T=w_j | d) =$  「 $d$  に含まれる単語  $w_j$  の数 /  $d$  に含まれる単語数」

$P(T=w_j) =$  「全文書に含まれる単語  $w_i$  の数 / 全文書に含まれる単語数」

$P(C_i) =$  「 $C_i$  に含まれる文書数 / 全文書数」

上記(6)式を用いると、各候補カテゴリ  $C_i$  について  $P(C_i | d)$  が計算出来る。処理機能1-4において、

$$P(C_i | d) = \sum_j [P(C_i | T=w_j) * P(T=w_j | d)]$$

( $\sum_j$  は全ての  $j$  に対する総和) (5)

となる。ここで、ベイズの定理を用いて  $P(C_i | T=w_j)$  を書きかえると、

計算した  $P(C_i | d)$  を降順にソートして、例えば、上位  $K$  個のカテゴリを選択し、文書  $d$  が持つカテゴリとする。

【0016】

【発明の効果】本発明で提案する文書の特徴付けによると、訓練用データの中に存在しないキーワードで特徴付けを行っても、データの補完をすることなしに文書自動分類が可能になる。

【図面の簡単な説明】

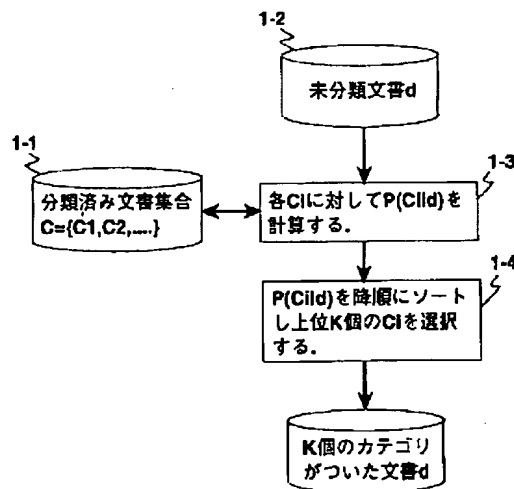
【図1】本発明の実施例の処理手順を示した図。

【符号の説明】

1-1：既に分類済みの文書集合、1-2：未分類の文書

【図1】

図1



## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-155758

(43)Date of publication of application : 06.06.2000

(51)Int.Cl.

G06F 17/30

(21)Application number : 10-328940

(71)Applicant : HITACHI LTD

(22)Date of filing : 19.11.1998

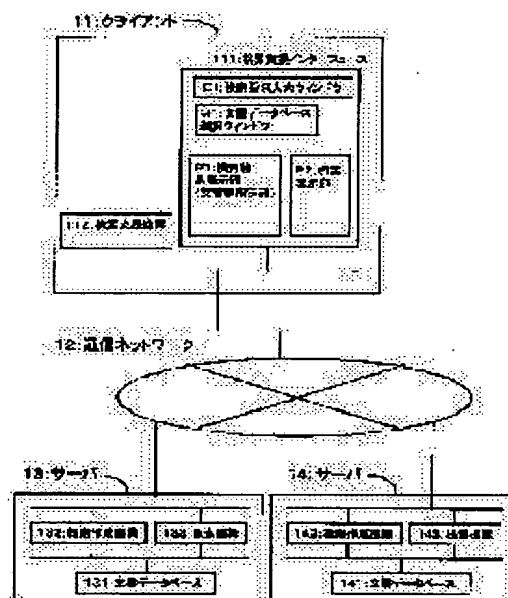
(72)Inventor : IWAYAMA MAKOTO  
NISHIOKA SHINGO  
NIWA YOSHIKI  
TAKANO AKIHIKO

## (54) METHOD AND SERVICE FOR DOCUMENT RETRIEVAL FROM PLURAL DOCUMENT DATA BASES

## (57)Abstract:

PROBLEM TO BE SOLVED: To efficiently realize document retrieval for checking the correlations between document data bases.

SOLUTION: The document data bases 131 and 141 are provided with summary generating mechanisms 132 and 142 and retrieving mechanisms 133 and 143 and are connected as servers 13 and 14 to a communication network 12. A client 11 obtains a relevant document group among document groups in a specified document data base through its summary. The obtained summary is sent to the other server to perform retrieval corresponding to the transferred summary from the document data base on the server at the transfer destination.



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2000 Japanese Patent Office